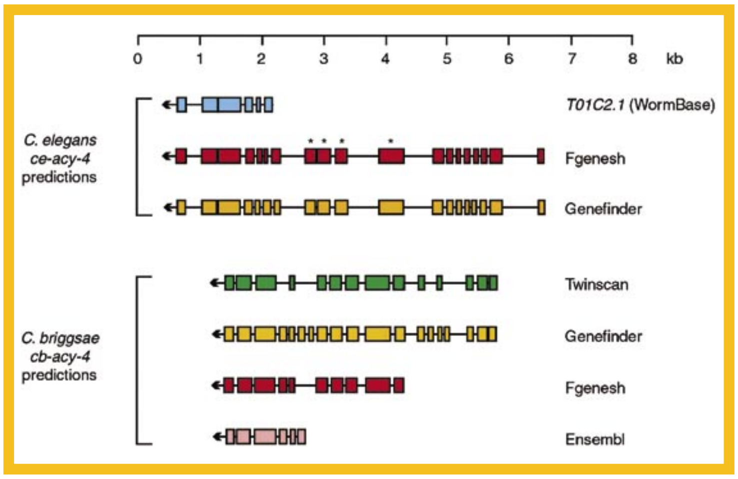
Research Digest

Synopses of Research Articles

Newly Sequenced Worm a Boon for Worm Biologists

Caenorhabditis elegans, a 1-mm soil-dwelling roundworm with 959 cells, may be the best-understood multicellular organism on the planet. As the most "pared-down" animal that shares essential features of human biology—from embryogenesis to aging—*C. elegans* is a favorite subject for studying how genes control these processes. The way these genes work in worms helps scientists understand how diseases like cancer and Alzheimer's develop in humans when genes malfunction. With the publication of a draft genome sequence of *C. elegans*' first cousin, *C. briggsae*, Lincoln Stein and colleagues have greatly enhanced biologists' ability to mine *C. elegans* for biological gold.

Every organism carries clues to its molecular operating system and evolutionary past embedded in the content and structure of its genome. To unearth these clues, scientists examine different regions of the genome, assembling data on sequences, genes, functional elements that are not genes (but that regulate them, for example), repeated sequences, and so on. By comparing the genomes of related organisms, researchers can see what parts of the genomes are conserved—highly conserved genes tend to be important—and then focus on these regions to track down genes and determine how they function.



Sequence comparison between the two worm genomes

To construct a draft sequence of the *C. briggsae* genome, the researchers merged genomic data from three sources—one derived from whole-genome shotgun sequencing, another from physical genome mapping, and the third from regions of a previously "finished" sequence. For the shotgun sequence, the researchers extracted DNA from worms, randomly cut it into short pieces, sequenced them, and then assembled overlapping sequences to create thousands of stretches of contiguous DNA sequence. To help fill in the gaps between these "contigs," Stein and colleagues developed a "fingerprint" map of the genome as a guide for aligning the shorter fragments. The map also helped them identify inconsistencies and misalignments in the genome assembly. Finally, they integrated the previously finished sequence to improve the draft genome sequence. Using these massive datasets, the authors produced a high-quality genome sequence; although it does not quite meet the gold standard of a "finished" sequence, it covers 98% of the genome and has an accuracy of 99.98%.

After confirming the accuracy of the draft, the researchers turned to the substance of the genome. Examining two species side by side, scientists can quickly spot genes and flag interesting regions for further investigation. Analyzing the organization of the two genomes, Stein et al. not only found strong evidence for roughly 1,300 new *C. elegans* genes, but also indications that certain regions could be "footprints of unknown functional elements." While both worms have roughly the same number of genes (about 19,000), the *C. briggsae* genome has more repeated sequences, making its genome slightly larger.

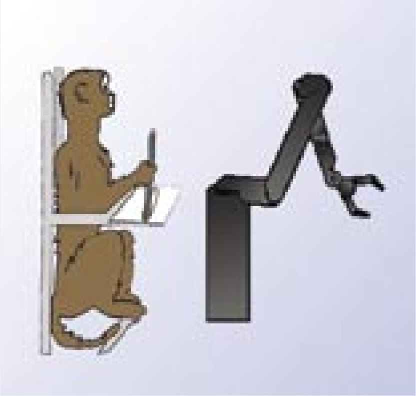
Because the worms set out on separate evolutionary paths about the same time mice and humans parted ways—about 100 million years ago, compared to 75 million years ago—the authors could compare how the two worm genomes have diverged with the divergence between mice and humans. The worms' genomes, it seems, are evolving faster than their mammalian counterparts, based on the change in the size of the protein families (*C. elegans* has more chemosensory proteins than *C. briggsae,* for example), the rate of chromosomal rearrangements, and the rate at which silent mutations (DNA changes with no functional effect) accumulate in the genome. This would be expected, the researchers point out, because generations per year are a better measure of evolutionary rate than years themselves. (Generations in worms are about three days; in mice, about three months.) What is surprising, they say, is that despite these genomic differences, the worms look nearly identical and occupy similar ecological niches; this is obviously not the case with humans and mice, which nevertheless have remarkably similar genomes. Both worm pairs—as well as mouse and human—also share similar developmental pathways, suggesting that these pathways may be controlled by a relatively small number of genes and that these genes and pathways have been conserved, not just between the worms, but also between the nematodes and mammals. This question, along with many others, can now be explored by searching the two species' genomes and comparing those elements that have been conserved with those that have changed.

With the nearly complete *C. briggsae* genome in hand, worm biologists have a powerful new research tool. By comparing the genetic makeup of the two species, *C. elegans* researchers can refine their knowledge of this tiny human stand-in, fill in gaps about gene identity and function, as well as illuminate those functional elements that are harder to find, and study the nature and path of genome evolution.

Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, et al. (2003) The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. DOI: 10.1371/journal.pbio.0000045

Retraining the Brain to Recover Movement

Some 200,000 people live with partial or nearly total permanent paralysis in the United States, with spinal cord injuries adding 11,000 new cases each year. Most research aimed at recovering motor function has focused on repairing damaged nerve fibers, which has succeeded in restoring limited movement in animal experiments. But regenerating nerves and restoring complex motor behavior in humans are far more difficult, prompting researchers to explore alternatives to spinal cord rehabilitation. One promising approach involves circumventing neuronal damage by establishing connections between healthy areas of the brain and virtual devices, called brain–machine interfaces (BMIs), programmed to transform neural impulses into signals that can control a robotic device. While experiments have shown that animals using these artificial actuators can learn to adjust their brain activity to move robot arms, many issues remain unresolved, including what type of brain signal would provide the most appropriate inputs to program these machines.



Monkey learns to control BMI

As they report in this paper, Miguel Nicolelis and colleagues have helped clarify some of the fundamental issues surrounding the programming and use of BMIs. Presenting results from a series of long-term studies in monkeys, they demonstrate that the same set of brain cells can control two distinct movements, the reaching and grasping of a robotic arm. This finding has important practical implications for spinal-cord patients—if different cells can perform the same functions, then surgeons have far more flexibility in how and where they can introduce electrodes or other functional enhancements into the brain. The researchers also show how monkeys learn to manipulate a robotic arm using a BMI. And they suggest how to compensate for delays and other limitations inherent in robotic devices to improve performance.

While other studies have focused on discrete areas of the brain—the primary motor cortex in one case and the parietal cortex in another—Nicolelis et al. targeted multiple areas in both regions to operate robotic devices, based on evidence indicating that neurons involved in motor control are found in many areas of the brain. The researchers gathered data on both brain signals and motor coordinates—such as hand position, velocity, and gripping force—to create multiple models for the BMI. They used different models according to which task the monkeys were learning—a reaching task, a hand-gripping task, and a reach-and-grasp task.

The BMI worked best, Nicolelis et al. show, when the programming models incorporated data recorded from large groups of neurons from both frontal and parietal brain regions, supporting the idea that each of these areas contains neurons directing multiple motor coordinates. When the researchers combined all the motor parameter models to optimize the control of the robotic arm through the BMI, they fixed those parameters and transferred control to the BMI and away from the monkeys' direct manipulation via a pole. The monkeys quickly learned that the robotic arm moved without their overt manipulations, and they periodically stopped moving their arms. Amazingly, when the researchers removed the pole, the monkeys were able to make the robotic arm reach and grasp without moving their own arms, though they did have visual feedback on the robotic arm's movements. Even more surprising, the monkeys' ability to manipulate the arm through "brain control" gradually improved over time.

One way the brain retains flexibility in responding to multiple tasks is through visual feedback. The researchers suggest that the success of the model may be the result of providing the monkeys with continuous feedback on their performance. This feedback may help integrate intention and action—including the action of the robotic arm—in the brain, allowing the monkey to get better at manipulating the robotic arm without moving.

By charting the relationship between neural signals and motor movements, Nicolelis et al. demonstrate how BMIs can work with healthy neural areas to reconfigure the brain's motor command neuronal elements and help restore intentional movement. These findings, they say, suggest that such artificial models of arm dynamics could one day be used to retrain the brain of a patient with paralysis, offering patients not only better control of prosthetic devices but the sense that these devices are truly an extension of themselves.

Carmena JM, Lebedev MA, Crist RE, O'Doherty JE, Santucci DM, Dimitrov DF, Patil PG, Henriquez CS, Nicolelis MAL (2003) Learning to control a brain–machine interface for reaching and grasping by primates. DOI: 10.1371/journal.pbio.0000042

Computer Model Predicts How the Brain Controls Limb Dynamics

If you have ever spent an evening hoisting brews with your pals at the corner pub, chances are you never stopped to think—gee, how do I lift my glass now that it's only half full? It seems like a simple task—you raise that glass reflexively, whether it is empty or full—yet the neural calculations that determine the force needed to lift your arm smoothly to your lips in each case are anything but simple.

The brain, it seems, operates like a computer to process variable cues—such as the weight of a glass and the position of your arm—to generate an appropriate response: lifting the glass. Neuroscientists believe the brain builds a kind of internal software program based on past experience to transform such variable cues into motor commands. The brain's software, or internal model, depends on specialized sets of instructions, or "computational elements," in the brain. But exactly how the brain organizes these elements to process sensory variables that affect arm movements is far from clear.

Eun Jung Hwang and colleagues predict that these computational elements are based on a multiplicative mechanism, called a *gain field*, through which sensory signals to the brain are amplified by signals from the eye, head, or limbs. In this way, the brain can rely on past experience of one kind of sensory cues to predict how to respond to new but similar situations. While previous studies had established that some visual cues are combined through a gain field, this study shows that motor commands may also be processed via gain fields. This finding, the researchers demonstrate, accounts for a range of behaviors.

Based on previous studies showing that when people reach to various directions in a small space, they can extrapolate what they learn about the forces in one starting position to a significantly different position, it has been proposed that the way the brain computes movement is not terribly sensitive to limb position. Citing other research with seemingly contrary conclusions—that the brain can be highly sensitive to limb position in calculating force and movement—Hwang et al. set out to investigate whether—and how—the brain creates a template to translate sensory variables (limb position and velocity) into motor commands (force). They created a computer model to mimic the reaching behaviors observed by people in their experiments and found that the most accurate model used computational elements that are indeed sensitive to both limb position and velocity. If the brain processes these two independent variables through a gain field, it can use the relationship of the two variables—that is, the strength of the gain field—to adapt information about the force needed to move or lift something in one situation to accomplish a wide range of similar movements. When the researchers compared their model to previously published results, they found their model accounted for seemingly disparate findings. They explain that the brain's sensitivity to limb position can be either low or high after a task has been learned because the gain field itself is adjustable.

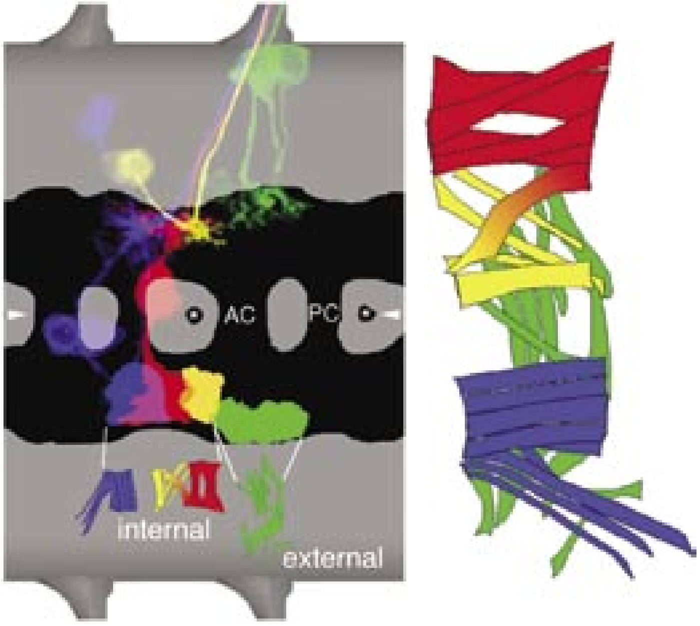
The authors note that neurophysiological experiments suggest that the motor cortex may be one of the crucial components of the brain's internal models of limb dynamics. The next step will be to track the motor cortex neurons to see whether their activity supports this model. Hwang et al. predict they will.

Hwang EJ, Donchin O, Smith MA, Shadmehr R (2003) A gain-field encoding of limb position and velocity in the internal model of arm dynamics. DOI:10.1371/journal.pbio.0000025

Underlying Principles of Motor System Organization Revealed

Time after time in biology, revelations about structure lead to insights about corresponding functional mechanisms. While evolution throws in the occasional spandrel, more often organizational structure serves a practical purpose. So naturally, neuroscientists wonder, does the architectural organization of the motor system reveal an underlying functional organization?

Progress on this question has been complicated by the fact that there appears to be no clear correspondence between the development of motor neurons centrally and their target muscles in the periphery. In the visual system, for example, retinal ganglion cells send axons in an ordered manner into the brain, where they form connections with neurons of the primary visual center in the brain responsible for detecting visual targets. The arrangement of these connections mirrors the neighboring relationships of the neurons in the retina, and so the neural map of connections in the brain is an "anatomical correlate" of the arrangements in the retina. The origin of these anatomical relationships can be traced through the process of development, allowing scientists to link the assembly of this sensory system with the function of the neurons involved. Matthias Landgraf and colleagues now report that in the fruitfly *Drosophila* the arrangement of motor neurons corresponds to the distribution of their target muscles. Thus, anatomical correlates also exist in the motor system, in the form of a "myotopic map," where the arrangement of motor neuron dendritic branches in the central nervous system reflects the distribution of their target body wall muscles in the periphery.



Organization of *Drosophila* motorneurons and their target muscles

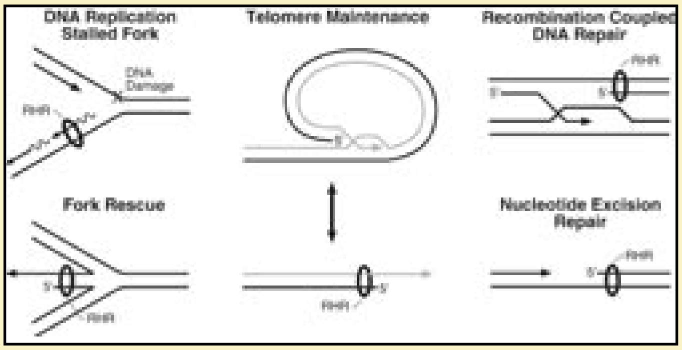
Starting with the larger question of how the neural networks governing locomotion are specified and assembled during development, the researchers decided to see if they could identify an elementary principle of motor system organization. Working in *Drosophila*, they examined motor neurons and the body wall muscles they innervate. With an eye toward understanding the mechanisms directing the assembly of the motor system, the researchers concentrated on the early stages of development, when the motor neurons first establish their characteristic dendritic territories. They found that the dendrites of motor neurons innervating internal muscles and that those innervating external muscles do in fact project into distinct regions, corresponding to the distinct mapping of the muscles themselves. Surprisingly, the arrangement of the dendrites in the myotopic map forms independently of the muscles they innervate. It may be, the researchers suggest, that the initial signals charting the location of the dendrites are set very early in development, when the coordinates for other structural elements are established. But that question requires further investigation.

The researchers are among the first to reveal such an orderly connection between patterns of motor neuron dendrites and patterns of muscles. This organization, in the form of the myotopic map, may be mirrored by the patterning of processes of higher-order neurons, which form connections with the motor neuron dendrites themselves. In vertebrates, studies have shown that motor neurons are grouped into "pools" and "columns" that correlate with the muscles they innervate. But because these pools and columns represent the location of the cell bodies and not the areas of the spinal cord where the neurons receive most of their inputs, that is, their dendritic branches, scientists could not say whether the pools and columns are simply spandrels—an incidental result of the way motor neurons are generated—or mirror a functional organization of the motor system. This novel finding in *Drosophila* will pave the way for future studies on the relationship between anatomy and physiology during development. It will be particularly interesting to discover whether such myotopic arrangements of motor neuron dendrites are unique to insects or whether this organizational principle occurs in other motor systems, including vertebrates.

Landgraf M, Jeffrey V, Fujioka M, Jaynes JB, Bate M (2003) Embryonic origins of a motor system: Motor dendrites form a myotopic map in *Drosophila*. DOI: 10.1371/journal.pbio.0000041

Novel "Checkpoint" Mechanism Mediates DNA Damage Responses

Of all the tasks a cell must accomplish day in and day out, protecting its genome may be the most important. Genomes confront all manner of potential assaults, from the strand-splitting action of gamma-radiation to the simple copying mistakes sometimes made when DNA replicates before a cell divides. Though some mutations are harmless, others can disrupt gene action, leading to cancer and other diseases. To guard against such events, healthy cells maintain quality-control "checkpoints" that sense and respond to DNA injuries, as well as to defects in DNA replication, and that prevent cell division until the DNA can be repaired. If the damage is beyond repair, apoptosis pathways set about the business of destroying the afflicted cell.


DNA Replication Stalled Fork
DNA Damage
RHR
Telomere Maintenance
Recombination Coupled DNA Repair
RHR
5'
5'
Fork Rescue
5'
RHR
RHR
5'
Nucleotide Excision Repair
RHR
5'

Possible targets for the RHR checkpoint clamp

Many of the genes and protein complexes involved in these checkpoint responses have been identified, but the biochemical mechanisms that in some cases trigger cell cycle arrest are not fully understood. Experiments by Philip Hanawalt and his student David Pettijohn at Stanford University in 1963 suggested that the molecular machinery of DNA replication and repair—which they discovered at sites of damage—are quite similar and closely linked. While many studies have since supported that link, Viola Ellison and Bruce Stillman, the director of the Cold Spring Harbor Laboratory, have found new evidence that the two processes may indeed coincide by showing that protein complexes regulating a cellular checkpoint in DNA repair operate much like similar complexes involved in DNA replication.

The molecular pathways governing the replication of DNA before cell division are well known. As the double-stranded DNA molecule unwinds, different protein complexes step in to ensure that each strand is faithfully reproduced. Two protein complexes required for this process are replication factor C (RFC) and proliferating cell nuclear antigen (PCNA). In the 1980s, Stillman's laboratory isolated PCNA and RFC and showed that they function together to "load" PCNA onto a structure in DNA that is created after DNA synthesis begins. PCNA forms a clamp around the DNA strand and regulates the DNA polymerases that duplicate the DNA double helix.

Studies in yeast had identified a series of proteins required for the DNA synthesis phase of the cell cycle and the DNA damage checkpoint pathways; mutations in these proteins' genes make cells very sensitive to radiation (hence the name *Rad* genes). A subset of these proteins, which are conserved in human cells, form two protein complexes—RSR and RHR—that function like RFC and PCNA, respectively, with RSR loading the RHR clamp onto DNA. Ellison and Stillman demonstrate that both pairs of "clamp-loading" complexes follow similar biochemical steps, but, significantly, RFC and RSR favor different DNA structures for clamp loading. While it was known that the RSR/RHR complexes exist in human cells, it had not been established that the two types of clamps prefer different DNA targets. The researchers also show that the RSR/RHR biochemistry depends on RPA, a protein known to be involved in the DNA damage-response pathway.

The discovery that RSR loads its RHR clamp onto a different DNA structure was unexpected; it suggests not only that the two clamp loaders have distinct replication and repair functions, but also how the checkpoint machinery might work to prevent DNA damage from being passed on to future generations. By establishing the chemical requirements of RSR/RHR interactions as well as the preferred DNA-binding substrate, the researchers have charted the way for determining the different functions of these cell cycle checkpoint complexes and how the complexes' different subunits affect these functions.

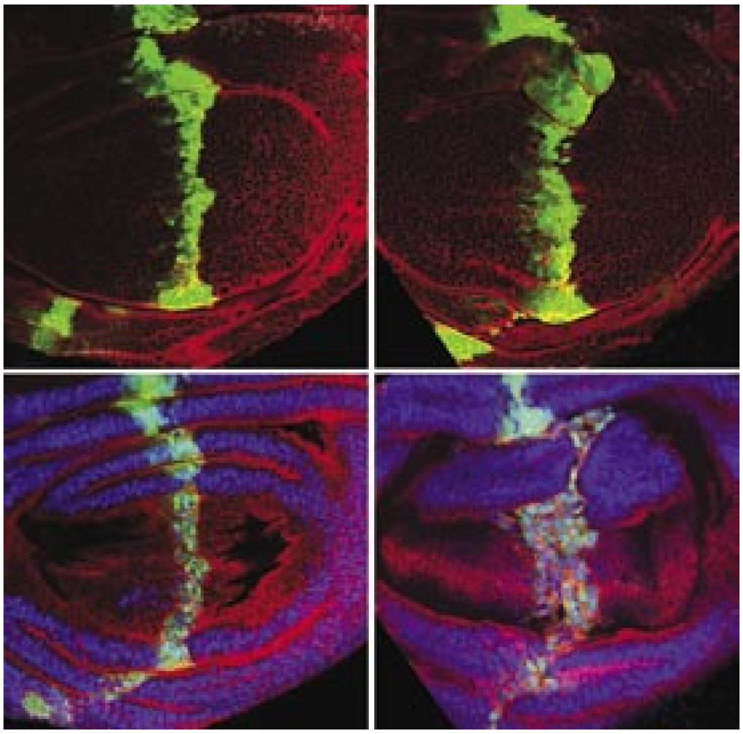
The researchers propose that the role of this checkpoint machinery is not as an initial sensor of DNA damage, but rather as a facilitator of DNA repair, stepping in after preliminary repairs to DNA lesions have been made. Ellison and Stillman's work helps establish a biochemical model for studying how both of these checkpoint complexes function to coordinate replication and repair—and promise to help scientists understand how cancer develops when the checkpoint repair mechanisms fail.

Ellison V, Stillman B (2003) Biochemical characterization of DNA damage checkpoint complexes: Clamp loader and clamp complexes with specificity for 5' recessed DNA. DOI: 10.1371/journal.pbio.0000033

slik Gene Controls Cell Growth and Survival

During animal development, cells gradually grow, multiply, and specialize to create the tissues and organs that shape and sustain multicellular organisms. The progression from a single cell to a thousand-, million-, or trillion-celled animal follows an exacting schedule and plan involving an elaborate network of genes and proteins. One of the primary mechanisms coordinating this process is cell-to-cell communication. Cellular signaling regulates two crucial development mechanisms, apoptosis (programmed cell death) and cell proliferation, which work like chisel and clay to sculpt multiplying masses of cells into, say, a fly wing or a human finger. Controlled by multiple signals operating at fixed intervals, the entwined pathways can be steered off-course by a single defect in the communication network, resulting in the death of a healthy cell, for example, or the survival of a damaged cell. Such disruptions can lead to physical abnormalities, such as webbed hands and feet, when cells that should die remain alive; degenerative nerve disease, when healthy cells are killed; and cancer, when damaged cells survive and evade normal growth limitations.

Researchers have uncovered some of the mechanisms underlying these processes by studying genes involved in fruitfly (*Drosophila*) development. Following that tradition, Stephen Cohen and David Hipfner have identified a gene critical to *Drosophila* development that juggles cell growth and survival signals to help promote cell growth and prevent inappropriate apoptosis. They searched for genes associated with changes in tissue growth in fruitfly wings and identified some that can cause tissue "overgrowth"—abnormally large masses resulting either from cells growing faster than they divide or from cells escaping proliferation controls when they are overexpressed. Among these is a gene that encodes a newly identified kinase that contributes to the regulation of cell proliferation and survival (or death, depending on the circumstance) during *Drosophila* development. Cohen and Hipfner called the gene *slik* based on its similarity to two human kinase-coding genes (*SLK* and *LOK*). Little is known about these human proteins, though previous studies suggest they may affect cytoskeletal dynamics and cell adhesion. In this paper, the authors report preliminary evidence supporting the notion that *slik* may regulate the cytoskeleton, the "backbone" of the cell that confers structure and motility. Interestingly, disturbances to cell adhesion and cytoskeletal structure are known triggers of apoptosis and are being explored as potential anticancer agents.



***slik* overexpression induces apoptosis**

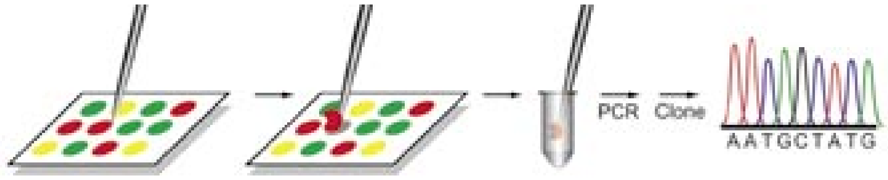
Kinases make up one of the largest families of proteins and are important regulators of cell signaling. To investigate the function of *slik* in *Drosophila*, the researchers removed the gene and then studied the physical and cellular effects. They found striking delays in growth and developmental timing and showed that these effects result largely from the demise of the *slik*-deficient cells. While cells deprived of *slik* can grow, divide, and differentiate, most respond to the defect by killing themselves, even under conditions that normally promote survival. Thus, cells without *slik* appear to have an intrinsic survival defect, suggesting that *slik* prevents apoptosis. When *slik* is overexpressed, cell proliferation increases, but so does apoptosis. Only when apoptosis was blocked did the cells form tumor-like growths. This coupling of cell growth and cell death is characteristic of oncogenes (cancer-causing genes), and *slik* also seems to function in both pathways. The authors point out that the signal to proliferate may inherently sensitize cells to apoptosis, as has been shown previously for some cancer cells. This may keep an individual cell under the control of its neighbors, who collectively monitor the needs of the organism. For a cell to respond to a signal by dividing rather than dying, it must get the appropriate signs from its comrades. *slik*, the authors demonstrate, is a key factor in determining whether a cell lives or dies. Whether its mammalian counterparts play a similar role is yet to be determined.

Hipfner DR, Cohen SM (2003) The Drosophila Sterile-20 kinase Slik controls cell proliferation and apoptosis during imaginal disc development. DOI: 10.1371/journal.pbio.0000035

Gene Chip for Viral Discovery

West Nile virus. Monkey pox. SARS. If the ever-growing list of public health scares has taught us anything, it's that we need quick, effective tools for detecting emerging viral threats. Researchers led by Joseph DeRisi of the University of California at San Francisco have combined genome databases of sequenced viruses with DNA microarray technology to create such tools.

The viral gene chip they created can



Extraction, amplification, and decoding of viral sequences

rapidly identify known viruses and classify new ones based on their genetic makeup. This was validated in March when the viral chip contributed to the identification of the cause for severe acute respiratory syndrome (SARS) as a novel coronavirus. In the article published in this issue, the researchers describe the chip (or microarray), how it was used in the classification of the SARS virus, and how it provides direct access to viral genomic sequence.

Microarray technology works by taking advantage of the structural properties of DNA. DNA molecules normally exist as double helices, two complementary strands of nucleotides wrapped around each other. The microarray consists of a large number of single DNA strands attached to a solid base. These probes (which in case of the viral chip represent sequences from all fully sequenced reference viruses) can be used to interrogate unknown sequences: if a solution containing such sequences is passed over the chip, similar sequences will "hybridize," or bond in a signature double helix.

Known viruses hybridize in a characteristic pattern and can be identified quickly. Because bonding occurs even when the match between probe and sample sequence is not perfect, new relatives of known viruses can be identified as belonging to a particular family (such as coronaviruses, in the case of SARS).

To quickly obtain more information on a novel virus, it is then possible to "syphon off" those viral sequences that stuck to their respective counterparts on the chip and to use the material to determine part of the genomic sequence. Such sequence information provides more detail on how the new virus relates to known ones, which might provide clues about its origin and possible treatment strategies.

A Single Protein in Yeast Can Fine-Tune an Environmental Response

One might not expect that yeast lead terribly eventful lives, yet the single-celled fungus must struggle to survive just like everyone else. And for yeast—like everyone else—survival means being able to detect and coordinate a rapid response to changes in its environment. Though survival for humans is a bit more complicated, our cells use the same regulatory networks, which maintain cell growth and health when they work and contribute to diseases, from asthma to cancer, when they break down.

Given the variety of conditions even the lowly yeast is likely to encounter during its life, one might expect to find a multitude of molecules mobilizing a response. But yeast cells, it turns out, are fairly resourceful. As Erin O'Shea and colleagues report, just one protein in yeast activates different groups of genes in response to different amounts of an environmental stimulus. The researchers focused on how yeast responds to various levels of phosphate, an essential nutrient for all cells.

One way that cells regulate responses to environmental stimuli is through the transcription (activation) of genes. These transcriptional responses are often controlled by a multistep process that shuttles gene-activating proteins into the nucleus, where they can generate the appropriate response for a given stimulus, or confines them to the cytoplasm if their gene products are not needed. During this process, called phosphorylation, the addition of a phosphate group to a protein—such as a receptor or transcription factor—acts as a mechanism for controlling gene expression.

O'Shea's team demonstrated that phosphorylation of a transcription factor called Pho4 controls gene expression by controlling where that protein resides in the cell—in the cytoplasm or in the nucleus. As is the case with many proteins, Pho4 can accept phosphate groups at multiple sites. To see whether the location of phosphorylation affects the action of Pho4, O'Shea's team exposed yeast to different levels of phosphate and tracked the cellular response.

They found that when yeast is deprived of phosphate, Pho4 has no phosphate groups at any of its binding sites and enters the nucleus, where it binds to DNA and activates a set of genes whose products can scavenge for phosphate or otherwise compensate for the scarcity. When yeast has ample supplies of phosphate, Pho4 is phosphorylated and remains in the cytoplasm—unable to influence transcription—suggesting that the cells can absorb plenty of nutrients from their environs without having to engage a specialized foraging team. When the researchers exposed the yeast to intermediate amounts of phosphate, the results were surprising. Middling concentrations of phosphate produced different forms of phosphorylated Pho4, which varied in their ability to activate genes, and so added to the number of possible responses. Pho4 partially phosphorylated at one site, for example, could still enter the nucleus, but activated only one type of phosphate-recovery gene and not others.

While it is not unexpected that differential phosphorylation could have different functional outcomes, the authors say, it is surprising that one enzyme acting on one transcription factor can create different phosphorylation patterns—and therefore different gene-expression patterns—in response to different amounts of a single stimulus. Their results show that cells rely on a highly regulated series of interactions that induce subtle changes in gene expression to fine-tune their response to small environmental changes. And they do this in a remarkably efficient manner, relying on a small cast of characters to orchestrate the responses essential for survival.

Springer M, Wykoff DD, Miller N, O'Shea EK (2003) Partially phosphorylated Pho4 activates transcription of a subset of phosphate-responsive genes. DOI: 10.1371/journal.pbio.0000028

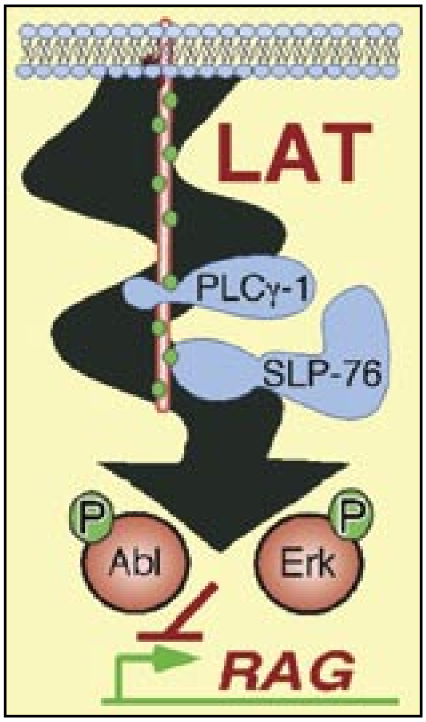
Basal Signaling Suppresses *RAG* Genes during T Cell Development

Faced with all manner of potential threats in the form of billions of different viral, bacterial, and chemical pathogens, the mammalian immune system relies on a "safety in diversity" strategy for protection. With two distinct subsystems—one innate, the other adaptive—the immune system can recognize some 100 trillion antigens. The innate system deploys cells programmed to quickly recognize microbes with a particular set of conserved molecular structures. The adaptive system relies on billions of uniquely outfitted lymphocytes (white blood cells) to identify just as many pathogens through their protein fragments, or antigens. A human being grinds out billions of these cells every day. In the absence of threats, the immune system maintains a quiescent state and many of these cells are discarded. But for the immune system, doing nothing takes a concerted effort.

Lymphocytes originate in the bone marrow, though not all differentiate there. One class of lymphocytes, called T cells, develops in the thymus, where every T cell acquires a one-of-a-kind receptor, called a T cell receptor (TCR), designed to recognize a different antigen. When an antigen gets bound by a TCR (a bound molecule is called a ligand), the antigen triggers a signaling cascade that tells the T cell either to attack the infected cell or to alert other immune cells of the infiltrator. But as Jeroen Roose, Arthur Weiss, and colleagues report, signaling pathways activated by bound TCRs appear to influence gene expression even in the absence of antigen or other receptor ligands, a process called ligand-independent signaling. These findings lend support to the notion that cellular signaling pathways regulated by surface receptors, like TCRs, exhibit a continuous low-level signaling (known as basal signaling) in the absence of a stimulus and that this continuous signaling, by influencing gene expression, has significant influence on cellular differentiation.

Roose, Weiss, et al. focused on the TCR signaling pathway that regulates the expression of a group of genes, including *RAG-1* and *RAG-2*, that are activated in two distinct waves during T cell development. *RAG* genes play a crucial role in T cell development, a highly complex, multistage process that involves a reshuffling, or recombination, of TCR genes and the activation of different proteins and genes at different stages. *RAG* genes regulate the genetic recombination and ultimate cell surface expression of TCRs. Using chemical inhibitors and mutant human T cell lines deficient in critical signaling components involved in antigen receptor-dependent pathways, the researchers found that the loss of specific functions or specific proteins affected an unexpected set of target genes. Notably, when downstream components (the protein kinases Erk and Abl) were disabled in the basal signaling pathway, the researchers saw a resurgence of *RAG* gene expression. While Erk was already known to play a prominent role in signaling pathways downstream of the TCR, it now appears that Abl may also be regulated in TCR pathways. Most importantly, these findings suggest that signaling pathways thought to be triggered only by ligated receptors can influence gene expression on their own. And it may be through this type of signaling that TCR pathways help regulate T cell development by repressing *RAG* gene activity.

These basal signals, the researchers postulate, may in effect save the *RAG* expression machinery until recombination is called for. If *RAG* genes were expressed at the wrong time, they could cause inappropriate genetic recombination and create T cells that either lack function or attack healthy cells, as happens in immunodeficiency and autoimmune diseases. Elucidating the mechanisms and components of this basal pathway will contribute important insights into the development and function of the immune system. But these studies also establish a model for investigating other signaling systems, to determine whether biologically functional basal signaling is a rare phenomenon or whether it is a fundamental cell process needed to control the profile of gene expression in the quiescent state.



Basal signaling suppresses *RAG* gene

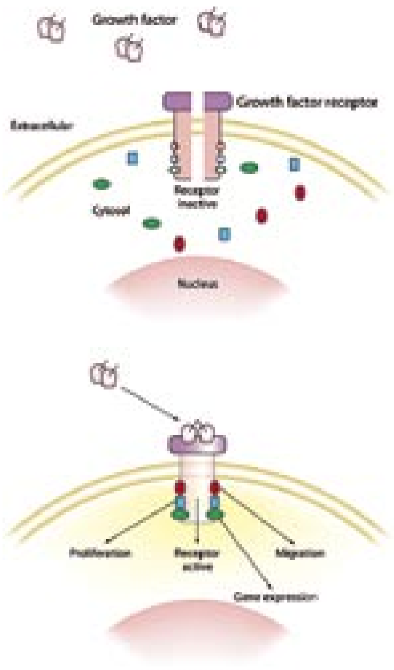
Roose JP, Diehn M, Tomlinson MG, Lin J, Alizadeh AA, et al. (2003) T cell receptor-independent basal signaling via Erk and Abl kinases suppresses RAG gene expression. DOI: 10.1371/journal.pbio.0000053

Development of Vascular Smooth Muscle Cells Depends on Signaling Synergy

A multicellular organism can have more than 200 different types of cells and as many as 100 trillion altogether. During the process of development, an organism enlists the service of hundreds of signaling molecules and thousands of receptors to direct cell growth, differentiation, and morphological destiny. Any given cell has no use for most of these signals and gets by with just a limited repertoire of receptors on its surface. Once a signal reaches a receptor, it triggers a series of biochemical reactions as different molecules transform the external signal into a biological response, in a process called signal transduction. One cell type controls all of its cellular functions—both universal and specialized—with just a few dozen receptors; each receptor elicits a wide range of responses by triggering a small number of interacting pathways. Exactly how a receptor produces the right response at the right time is a fundamental question in biology.

Of particular interest is a class of receptors—called receptor tyrosine kinases (RTKs)—that regulate cell proliferation, differentiation, and survival and play an important role in embryonic development and disease. Growth factor receptors are an important subset of RTKs. The platelet-derived growth factor receptor (PDGFR) family activates downstream signaling enzymes that stimulate the growth and motility of connective tissue cells, such as vascular smooth muscle cells (VSMCs), oligodendrocytes (cells of the tissue encasing nerve fibers), and chondrocytes (cartilage cells). The PDGF beta receptor is essential for directing the differentiation of VSMCs. While studies of signal transduction of this growth factor have established a model of how receptor tyrosine kinases function, the role of individual downstream signaling components in a living organism is still unclear.

Using mouse molecular genetics, Michelle Tallquist and colleagues set out to determine the function of individual components in the PDGFR beta pathway. They discovered a quantitative correlation between the overall amount of signal produced by the receptor and the end product of the signal, formation of VSMCs. Receptor responses, they report, are controlled in two ways: signaling was influenced both by the amount of receptors expressed and by the number of specific pathways engaged downstream of the receptor.



Model of growth factor receptor

Surface receptors have "tails" that project into a cell's interior. When a surface receptor is activated, a number of potential binding sites—modified amino acid residues—are exposed on its intracellular tail. Ten of these sites can bind to proteins with a specific amino acid sequence, called an SH2 domain; proteins with these domains can then initiate a signal transduction pathway. By introducing mutations in the SH2 domain-binding sites in mice, the researchers could evaluate how the loss of a particular binding site—and therefore pathway—affected the function of the receptor. They had previously investigated the functions of two other downstream signaling proteins in similar experiments.

Surprisingly, Tallquist et al. found that losing some of the individual components did not produce a significant negative physiological effect. Only when multiple downstream signaling pathways were disrupted did the researchers see a significant effect on the population of the cells. Reductions in the numbers of both activated receptors and activated signal transduction pathways produced reductions in the population of VSMCs. These results have not been seen in tissue culture before, suggesting that signal transduction is more complex in vivo and that future studies would benefit from incorporating a global approach, rather than targeting a single signaling component. The next step will be to investigate exactly how the individual pathways contribute to this result. It is also unclear whether these results apply only to these growth factor receptors or explain how RTKs operate in general.

Such questions have significant clinical relevance. Overexpression of the PDGFR beta pathway has been linked to a variety of serious diseases, including atherosclerosis and cancer. Understanding how cells control the action of this growth factor is an important step in developing targeted therapies. Since many of these conditions result from a growth factor stuck in the "on" position, inhibiting overactive receptors promises to be an effective clinical intervention. ■

Tallquist MD, French WJ, Soriano P (2003) Additive effects of PDGF receptor β signaling pathways in vascular smooth muscle cell development. DOI: 10.1371/journal.pbio.0000052

Essay

Genomics Research and Malaria Control: Great Expectations

Vincent P. Alibu, Thomas G. Egwang

The protozoan parasite *Plasmodium falciparum* causes falciparum malaria, a fatal parasitic disease in humans, and is transmitted by *Anopheles* mosquito vectors (predominantly the *Anopheles gambiae* complex and *An. funestus* in Africa). There are about 300 million malaria cases and 1–2 million deaths annually, the brunt of which are borne mostly in Africa by children under 5 years of age and by pregnant women. In many African countries, malaria poses a formidable challenge to an overburdened and underfunded public health system. The current malarial control strategies consist of chemotherapy directed against the malaria parasite and prevention of mosquito vector/human contact using insecticide-impregnated bednets and, to a lesser extent, indoor residual insecticide spraying and environmental control for reducing mosquito breeding sites. There are still no malaria vaccines in clinical practice.

The Dual Problem of Drug and Insecticide Resistance

Chemotherapy (the use of drugs to target disease) is used for both treatment and prevention. Drug resistance is increasingly becoming a problem. Some of the antimalarial drugs in current use include quinolines, artemisinins, antifolates, atovaquone/proguanil, and antibiotics. Chloroquine

Vincent P. Alibu is at the Zentrum für Moleculare Biologie at the University of Heidelberg in Germany. Thomas G. Egwang, an International Research Scholar of the Howard Hughes Medical Institute, is in the Department of Medical Parasitology at the Medical Biotechnology Laboratories in Kampala, Uganda. E-mail: egwang@imul.com

DOI: 10.1371/journal.pbio.0000039

(CQ) is a cheap and widely used aminoquinoline, but CQ-resistant parasites have become ubiquitous in endemic countries and other drugs are now used much more frequently (Ridley 2002). Fansidar, a combination of sulphadoxine and pyrimethamine (SP), is a first-line treatment in several African countries, but resistance to SP is spreading rapidly. Targeting the mosquito vector with pyrethroid-impregnated bednets, in addition to chemotherapy, is an effective method of controlling malaria transmission. However, pyrethroid resistance has been reported in *An. gambiae* s.s. in West Africa, and there is concern about its emergence in East Africa (Chandre et al. 1999). Thus, the public health problem due to malaria is exacerbated by the emergence of drug-resistant parasites and insecticide-resistant mosquitoes. The clinical application of efficacious intervention tools is therefore an urgent imperative for malarial control. This brings into sharp focus the importance of genomics research for drugs, vaccines, diagnostics, and insecticides. The unraveling of the genomes of humans, *P. falciparum*, and *An. gambiae* has ushered in a new era of hope that genomics research will result in the development of new and better tools for malaria control.

Early Pickings from the Malaria Genome

The *P. falciparum* genome of 22.8 megabases (Mbp) distributed among 14 chromosomes consists of 5,300 protein-coding genes (Gardner et al. 2002). *P. falciparum* possesses a relict plastid, the apicoplast, homologous to the chloroplasts of plants and algae. The apicoplast is essential for parasite survival and functions in the anabolic synthesis of fatty acids, isoprenoids, and heme (Seeber 2003). These essential metabolic pathways are not present in humans and are therefore ideal targets for the development of safe antimalarial drugs. Inhibitors of type II fatty acid biosynthesis (triclosan and thiolactomycin) and mevalonate-independent isoprenoid biosynthesis (fosmidomycin and FR900098) with potent antimalarial activities have been identified by computational mining of the genome data. The fact that fosmidomycin has rapidly entered into clinical trials underscores the great utility of genomics research in the control of malaria (Lell et al. 2003).

Malaria Functional Genomics

About 3,200 proteins (60%) in *P. falciparum* have no known functions (Gardner et al. 2002). The greatest challenge of malarial functional genomics (the elucidation of the functions of genes encoded by an organism's genome) is to assign functions to these proteins, thus comprehensively identifying the proteins that function at various lifecycle stages and that function together to carry out particular cellular processes, e.g., red blood cell invasion, signal transduction, growth, vesicular trafficking, etc. The application of functional genomics approaches allows the properties of many genes and proteins to be assessed in parallel on a large scale. These approaches are being used to address specific questions about the biology of *P. falciparum*. Gene profiling (determining which genes are expressed) by microarray technology allows a rapid, parallel analysis of genome-wide changes in gene expression over a variety of experimental conditions (e.g., chloroquine versus saline control), tissues, and cell types; these genes can be clustered (ordered by expression pattern) to identify those that function in the same process. One of the most promising applications of microarrays is the study of differential gene expression during the complex *P. falciparum* lifecycle, specifically the formidable and challenging task of determining which subset of the 5,300 genes is represented in the transcriptome of each stage (Bozdech et al. 2003; Le Roch et al. 2003). These approaches are beginning to yield invaluable insights about new vaccine candidates, novel drug targets, and the molecular basis of drug resistance.

Proteomics is the study of all the proteins expressed in an organism. Global protein analysis offers a unique means of determining not only protein expression, but also interacting partners, subcellular localizations, and post-translational modifications of proteins of whole proteomes. Analyses of the proteomes of parasites that have been exposed to distinct environmental stimuli (e.g., chloroquine versus saline control) or that manifest distinct phenotypes (drug resistant versus drug sensitive) might also facilitate the identification of biochemical drug targets and of the specific proteins involved in drug resistance. Comparative genomics (the comparison of genomes of related species), on the other hand, will yield invaluable insights about the biology of and the pathogenesis of disease associated with different parasites, i.e., *P. falciparum*

DOI: 10.1371/journal.pbio.0000039.g001

Figure 1. Patients at Apac Hospital
Of the patients waiting at the Out-Patient Department of Apac Hospital in Northern Uganda, the majority are mothers of children under 5 years old with malaria. (Photograph by Toshihiro Horii, Department of Molecular Protozoology, Research Institute for Microbial Diseases, University of Osaka, Osaka, Japan.)

on the one hand and *P. vivax* on the other. The biology and pathology of the two parasites are quite distinct, e.g., the preference for reticulocytes (*P. vivax*) versus mature red blood cells (*P. falciparum*), the ability to cause severe (*P. falciparum*) versus mild (*P. vivax*) disease, and the implication of amino acid substitutions in PfCRT in CQ resistance in one (*P. falciparum*) but not in the other (*P. vivax*).

The Promise of Mosquito Genomics

The 278 Mbp sequence of the nuclear genome of the PEST strain of *An. gambiae* s.s. has been published in draft form and is considerably larger than the 122 Mbp assembled sequence of the fruitfly *Drosophila melanogaster* (Holt 2002). The *An. gambiae* genome includes a treasure trove of 79 odorant receptor genes and about 200 genes that encode glutathione-S-tranferases, cytochrome P450s, and carboxylesterases. These and possibly other genes probably play a critical role in human host finding and detoxification of insecticides, respectively, and could be exploited, using gene profiling, proteomics, and comparative genomics, for the development of novel mosquito repellants or traps and insecticides. The ability to introduce foreign genes into *Anopheles* vectors is an exciting advance that might facilitate the development of transgenic mosquitoes that do not transmit malaria parasites (Moreira et al. 2002). However, the future implementation of this control strategy, if current technical hurdles can be overcome, must take into consideration concerns about the environmental impact of releasing genetically altered mosquitoes.

Capacity Building in Endemic Countries

Scientists in endemic countries must be active participants in malaria genomics research and not just conduits for field materials for Northern partners. However, the reality is that there is an increasing technological gap between endemic- and developed-country researchers in the field. This needs to be urgently addressed. The World Health Organization Special Programme for Research and Training in Tropical Diseases have initiated a series of training workshops in bioinformatics in endemic countries; the Howard Hughes Medical Institute has supported one such workshop. The training must extend to other aspects of genomics and include infrastructure development.

Great Expectations

There is considerable optimism that genomics research will result in new drugs, vaccines, diagnostics, and tools for malarial vector control. Strong linkages between genomics research and national malarial control programs will facilitate the translation of research findings into intervention tools. As it is for all new technologies, it might also be important for the communities in endemic countries to have a greater awareness and understanding of genomics research. This will enhance acceptance of the products and improve informed consent. There is therefore a unique opportunity for collaborations between social-economic scientists and genomics researchers. ■

Acknowledgments

We are grateful to Drs. Lisa Ranford-Cartwright (University of Glasgow, Glasgow, United Kingdom), Ayoade Oduola and Yeya Toure (World Health Organization, Geneva, Switzerland), and Wilfred Mbacham (University of Yaounde, Yaounde, Cameroon) for critical comments on the manuscript.

References

Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, et al. (2003) The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol 1: 10.1371/journal.pbio.0000005.

Chandre F, Darrier F, Manga L, Akogbeto M, Faye O, et al. (1999) Status of pyrethroid resistance in *Anopheles gambiae sensu lato*. Bull WHO 11: 230–234.

Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419: 498–511.

Holt RA (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298: 129–149.

Lell B, Ruangweerayut R, Wiesner J, Missinou MA, Schindler A, et al. (2003) Fosmidomycin, a novel chemotherapeutic agent for malaria. Antimicrob Agents Chemother 47: 735–738.

Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, et al. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 31 July 2003: e-publication ahead of print.

Moreira LA, Ghosh AK, Abraham EG, Jacobs-Lorena M (2002) Genetic transformation of mosquitoes: A quest for malaria control. Int J Parasitol. 32: 1599–1605.

Ridley RG (2002) Medical need, scientific opportunity and the drive for antimalarial drugs. Nature 415: 686–693.

Seeber F (2003) Biosynthetic pathways of plastid-derived organelles as potential drug targets against parasitic apicomplexa. Curr Immun Endocr Metabol Disord 2: 99–109.

Feature

Tough Mining

The challenges of searching the scientific literature

Steven Dickman

The standard "front end" for biomedical literature search is MEDLINE and its Entrez query system. Huge, well-managed, and nearly exhaustive, MEDLINE and its 11 million references provide incredible ease and facility for anyone who can type a Boolean query. Though not quite a parallel for Google—which runs a kind of popularity contest for Web links in real time—the Entrez search has opened up the literature to anyone with a Web browser. To those who grew up chasing citations and papers through the aisles of a scientific library, Entrez is a dream come true.

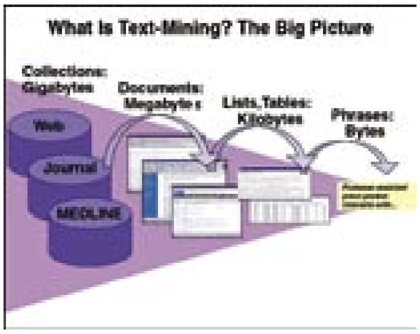
And yet. Suspend disbelief and imagine for a moment a kind of literature search dream-tool. "Find me all references citing my gene of interest," you could ask. But why stop there? "Find me all references citing some or all of my four genes of interest with expression or in vitro data." And then, "Bring up the text of the paragraph in which these citations occurred so I can view them in context. And do it in real time."

Tools that can perform such searches would go beyond Google because they avoid the repetitiveness involved in multiple searches. And they would go beyond Entrez because they would search the entire medical literature in full-text format and not, as MEDLINE does, just the abstracts. Furthermore, they would go beyond both types of searches in that they would be at least somewhat intelligent.

Such text-mining efforts are the next frontier for both academic and commercial groups that have sprung up from Pasadena to Boston to Tel Aviv.

Steven Dickman is a freelance writer and president of CBT Advisors in Cambridge, Massachusetts, United States of America. E-mail: sdickman@cbtadvisors.com

DOI: 10.1371/journal.pbio.0000048



DOI: 10.1371/journal.pbio.0000048.g001

Figure 1. Barely Getting below the Surface
The four levels of information retrieval: Google and MEDLINE both use keywords to direct a searcher to documents. But the next level has been tough to crack. Improved software would allow biologists to jump from the Web or MEDLINE to specifics with a single query. (Adapted with permission from the MITRE Corporation.)

But how realistic is this venture? Text-mining and its more universal relative "information retrieval" are still in their infancy. The first paper on text-mining for biology was published only in 1997. Furthermore, because biological text-mining comes so close to the challenge of comprehending human language—arguably the most complex invention in the history of the planet—it is what computer scientists call a "hard problem." So even here, at the embryonic and fun stage in this technology's history, the outcome and especially the timing of improvement are impossible to predict.

Build or Buy?

Language-processing software tools have been successfully applied in text-mining of nonscientific sources, especially to newswire content. Computer programs can already perform all three levels of text-mining (Figure 1) effectively: *retrieving* documents relevant to a given subject; *extracting* lists of entities or relationships among entities; and *answering* questions about the material, delivering specific facts in response to natural-language queries.

Information retrieval and extraction can be performed on news data at success rates of 90%–95%, says Lynette Hirschman, a structural linguist. Question-answering has been reported in the literature at 85% accuracy, she notes, which is "amazingly good." The question is, how soon can these levels be achieved for biology?

Good thing for biologists that Hirschman has turned her energies in their direction. Hirschman works in Massachusetts at MITRE Corporation, a government-funded institution that pursues projects in the national interest, be they in defense and intelligence or, as in the case of text-mining, "anywhere we can move an entire field forward," says Hirschman.

The good news from news-mining is that improvement seems to arrive in direct proportion to the time and energy expended by the research community. Similar improvement has occurred in speech recognition by computers, she adds (Figure 2). When people took successively harder problems and worked on them for four or five years, she explains, it caused error rates to drop, as a rule, by a factor of two every two years.

One might think tackling the biomedical literature would be relatively easy, remarks Hirschman: biology jargon has a lot of prefixes and suffixes, which can be parsed more easily than verbs and adverbs; it is highly regular, with Greek-letter add-ons to gene or protein names signifying relatives or subtypes of the original proteins; and there are many resources available, such as databases and ontologies linking different biological terms.

> *Information retrieval and extraction can be performed on news data at success rates of 90%–95%. The question is, how soon can these levels be achieved for biology?*

But whereas extraction of person and place names from news text routinely reaches 93%, results in biology remain mired in the 75%–80% range. "It's a little depressing," warns Hirschman. "Even something as simple as a slash may imply two different entities or a single compound."

A chorus of assent greets her observation. Programmers eager to codify the rules of biology have been stymied by what one bioinformaticist calls "a sea of exceptions."

Moreover, there is a chronic lack of data that have been "marked up" by software or humans to indicate the roles played by some of the key words. This marking-up process, however it is done, is crucial for machine-learning tasks. Getting these data is both hard and expensive, says Hirschman. To move biology text-mining forward, she believes, requires organizing different academic and commercial groups so that they are at least working on the same problem. Only then can standards emerge that will allow progress in the field even to be measured.

This type of shared problem—known as a "challenge evaluation"—has become something of a "religion" in the speech and language community since the 1980s, says Hirschman. By putting out a set of data to train on and then issuing a "challenge" for each group to extract the same information or answer the same questions, "you compare apples to apples. In the process you build a research community."

Last year, Hirschman and others ran the very first challenge evaluation in biology, the KDD Cup (officially called the Knowledge Discovery and Data-Mining Challenge Cup). Six weeks in advance, the organizers gave participants a training set of 862 journal articles already included in the model organism database FlyBase, along with associated lists of genes and gene products, as well as relevant data fields from FlyBase. After building their software tools, the entrants were then asked to take a test set of 213 articles and pretend they were curators: the tools were supposed to determine whether the articles were appropriate for curation, based on whether they contained experimental evidence for gene expression products, including both RNA transcripts and proteins.

Eighteen participants took a shot at the KDD Cup and their results speak of the infant state of the field. On average, they could assign only 58% of the papers correctly and could determine whether relevant gene products were present only 35% of the time. The winning entrant, a joint group from the Israeli company Clearforest ("see the forest *and* the trees") and Maryland-based Celera Genomics, did better.

Their entry made the right decision to curate 78% of the time and the right call on the presence of gene products 67% of the time.

The winning group did so well by using a clever "trick," says Hirschman admiringly. Their program searched for figure captions and then applied multiple techniques to find those gene products they were looking for.

Getting Out of the "Bag of Words"

The techniques applied by Clearforest and others fall into two broad categories, statistical and heuristic. Statistical techniques are the next step up from keyword searches. They count words such as genes or gene products appearing close to one another, but apply no linguistic insights, such as whether an adjective modifies a noun. By contrast, heuristic approaches use hand-crafted rules designed for specific datasets: e.g., January, February, March, etc., are months; the word following "Mr." is a name; and so forth. This approach is labor-intensive but especially useful when there is only a limited amount of data—as is the case with single scientific papers or small groups of papers.

Some statistical approaches have been labeled with the nickname "bag of words" because they fail to account for grammatical relationships; e.g., "man bites dog" and "dog bites man" would drop the same three words in the bag. A key observation at the KDD Cup was that the most basic statistical approach, which counts word occurrences at the document level, is not sufficient unless it takes into account at least some higher-level context, such as the part of the paper from which the search terms are extracted.

Furthermore, the more hand-crafted rules there were, the better. Many of the top teams included biologists who applied their expertise to help create empirical rules that became part of the program instructions. This points to a general theme in machine learning: the greater the degree of human intervention, the better. The best programs are covered with fingerprints.

Access: A Wrench in the Works?

Although the march toward better text-mining systems is building momentum, there are two issues that could stop it in its tracks. The first is access.

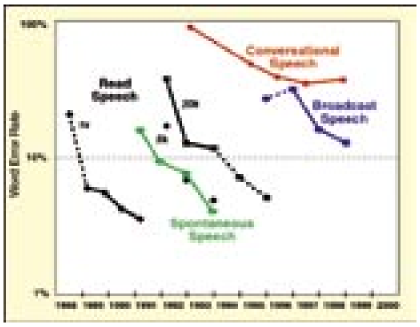
Experts in text-searching uniformly cite access as a key obstacle for developing better search tools. "Access is a bigger problem than algorithms" is how one machine-learning expert puts it, and a half-dozen others agreed.

The present "balkanized" situation for text-processing is filled with "dead ends" and "short circuits" in information flow among biologists, says David Lipman, head of the United States' National Center for Biotechnology Information, which runs PubMed, the MEDLINE database, as well as the National Library of Medicine and other critical resources in biology and bioinformatics. It is as if readers are marine biologists on a coastline whose beaches are 98% private. At best, asking permission to view every article slows down the work. At worst, there are some important tools one can never build owing to the missing context. MEDLINE itself would be much more powerful if it were based on full text, experts say.

Owing to lack of access, says Hirschman, "we miss a great deal by not having large corpora of full-text articles" included in the design of both the KDD Cup and the next challenge evaluation, called BioCreative, being held later this year. Many of the relevant biological data are found outside abstracts, but getting access to full text is complicated at best. For manual searching, researchers traditionally fall back on portal-hopping: jumping from one full-text subscription (to *Nature*, *Science*, or *Cell*, for example) to another, or from one portal (HighWire, Web of Science) to another. That way, many scientists routinely obtain access to as many as 80% of the journals they need. The rest they can usually request via interlibrary loan or order as photocopies online. However, this approach fails for most automated search programs. Just sorting out the permissions and keeping up with changes in the portals dramatically increase the headaches for anyone trying to build a search tool.

Laying Heisenberg to Rest

The second threat to text-searching programs ever becoming widely useful has more of the ring of linguistics jargon. The so-called " ontology problem" threatens successful searching based on the very specific nature of biological terminology.



DOI: 10.1371/journal.pbio.0000048.g002

Figure 2. The Impact of Challenge Evaluations (and Investment Dollars)
Driven by investment and competition as well as the pressure of regular challenge evaluations, error rates in speech recognition have dropped steadily, to the point where the technology has become standard from directory assistance to travel to financial information. Error rates drop by a factor of two every two years as challenge evaluations attract wide participation. (Graph adapted with permission from the MITRE Corporation.) Source: Pallett D, Garofolo J, Fiscus J (2000) Measurements in support of research accomplishments. Communications of the ACM: Special section on broadcast news understanding.

The issue here is not only that scientists are truly terrible about sticking to established terminologies. "Scientists would rather share each other's underwear than use each other's nomenclature," as biochemist Keith Yamamoto is fond of saying. Consequently, the scientific literature is a hodgepodge of identical or overlapping terms. A naïve text-parsing program does not know whether "cat" refers to the catalase gene, the chloramphenicol transferase gene, or a household animal.

The challenge is to build an ontology describing all the important relationships so your computer program can navigate among them without asking you what to do. Consequently, an ontology would prescribe rules for understanding the interactions among genes based on the appearance of certain verbs ("inhibit," "express"), nouns ("agonist"), or phrases. Although within each narrow scientific subdiscipline it may be possible to build exquisitely useful text-mining tools, as soon as programmers broach the borders of the narrowest subfields, they will run into a kind of Heisenberg uncertainty principle of linguistics and science. Every toolmaker is faced with the ontology problem in one respect or another, especially when the tool is meant to be a general one.

David Gilmour, chief executive officer of Tacit Inc., a knowledge management company in Palo Alto, California, is an industry veteran of exactly this war "and I have scars all over my body to prove it," he says. The issue in a nutshell, he explains, is that "ontologies scale poorly, and by the time they are useful," that is, large enough to capture most of the possible relationships among words, "they are unmaintainable."

Hirschman acknowledges that keeping up with the literature and new terminologies is challenging. Adapting tools to new domains has traditionally been one of the "critical stumbling blocks" for text-processing technology, she says. The dynamic growth of biological terminology does not help. There are 50–100 alterations *every week* to the nomenclature section of mouse genome database Web page.

Textpresso, Anyone?

Staying within one's narrow domain, then, could be a recipe for success, as long as the vocabulary and user questions remain tightly constrained, especially if there is a way to tiptoe around the access problem. That is apparently the case at Wormbase, though the newly available tool there, called Textpresso, is still being built. The motivation for Textpresso was simple, says Hans-Michael Mueller, a postdoctoral fellow in the lab of Paul Sternberg at Caltech in Pasadena, California, where Wormbase—the genetic database for the nematode worm *Caenorhabditis elegans*—is curated. "We want the user to be able to avoid going to the library to read all those papers [on genes and proteins] that your favorite gene interacts with. That is very tedious." The other goal is equally recognizable in the biology community: no mere mortal can hope to keep up with the burgeoning literature, even in the relatively narrow field of worm biology.

Mueller, a nuclear physicist by background, called Textpresso "a search engine for full-text searches of abstracts and articles" that can help find answers to more challenging queries than simple keyword searches.

Mueller and his team use human "taggers" to mark up the corpus of text to indicate categories like "biological processes" ("late larval activation"), "genes" (let-7), and "molecular functions." Then, like the Clearforest-Celera program, Textpresso searches for combinations of categories in the same or neighboring sentences. The ontology relating the expressions and categories to one another is based both on scientific and common sense as well as linguistic components. In less than two years of work, Mueller and his team have already marked up 3.9 million terms in 16,000 abstracts and 2,000 full-text papers. A typical search asks a question such as "What can be found out about the negative regulatory aspects of a genetic network in the pharynx?" Answers emerge in the form of citations, abstracts, and, if available, a paragraph or so from the text of the relevant paper. Textpresso went up—unpublicized—on the Web in February this year and already receives a couple of hundred hits a day, a big number in a field of about 2,000 researchers. Mueller estimates that Textpresso is 95% accurate and that about 35% of the relevant papers have been included.

Textpresso needs full-text access to be as good as it is, says Mueller. "We noticed" that drawing on full text "greatly increased the chances of a true hit," not a false positive. He managed to avoid the access issue by claiming a kind of "curator's privilege." Only the curators see the full text. Once the data are on the Web, users can only get at most a paragraph, which falls within fair use, said Mueller. If a user happens to subscribe to the journal in question, it is possible for him to click through, the publisher's portal and see the paper.

Whereas Textpresso works exclusively on worm genetic data and commercial players like Clearforest are just beginning to hunt for biological applications, a handful of companies have begun to market text-searching products to academic biomedical scientists. One such product is called QUOSA, for query, organize, share, and analyze. The software had its commercial launch in late 2002. Put simply, the program—available on an institution-wide basis and already installed for hundreds of researchers at Massachusetts General Hospital and the Dana-Farber Cancer Institute in Boston—allows a search across one's own documents. A front end for the literature that cooperates with MEDLINE, QUOSA pulls in and prioritizes full-text papers. The program first allows the user to search for the relevant files and download them in full-text format to the extent permitted by her library's subscription agreements and licenses. Once it becomes second nature to users, they rave about it. Like the best of the first-generation software, QUOSA allows users to make connections they would not have otherwise made. Like so many other early software products, its long-term success will hinge on demand as well as improvements made in the upgrades.

Because of the ontology problem, improvements in searching in the next couple of years are likely to result from the application of ever-better techniques within existing domains. Collaborations among Wormbase, Flybase, and other model-organism database groups will help improve all their search tools. MEDLINE itself may benefit from more advanced search techniques, though these will be restricted to abstract searches.

The big unknown for predicting further development of text-search tools is the path publishers will take. If each publisher or portal such as Reed-Elsevier or HighWire were to license or develop its own tool for searching its own content, the result might be better than the status quo, but would still be unsatisfying. Running the same search three times on three different subsets of content might be better than running it 15 times—but wouldn't it be easier to run it just once? ■